Serial production line performance under random variation

Romero Silva, R.; Marsillac, Erika; Shaaban, Sabry; Hurtado, Margarita

published in
Journal of Manufacturing Systems
2019

DOI (link to publisher)
10.1016/j.jmsy.2019.01.005

document version
Publisher's PDF, also known as Version of record

Journal of Manufacturing Systems 50 (2019) 278–289

Contents lists available at ScienceDirect

Journal of Manufacturing Systems

journal homepage: www.elsevier.com/locate/jmansys

Serial production line performance under random variation: Dealing with the 'Law of Variability'

Rodrigo Romero-Silva[a,b,*], Erika Marsillac[c], Sabry Shaaban[d], Margarita Hurtado-Hernández[a]

[a] Faculty of Engineering, Universidad Panamericana, Augusto Rodin 498, Insurgentes Mixcoac, 03920, Mexico City, Mexico
[b] Aviation Academy, Amsterdam University of Applied Sciences, Weesperzijde 190, 1097 DZ, Amsterdam, the Netherlands
[c] Information Technology, Decision Sciences and Maritime Supply Chain Management Department, Old Dominion University, 5115 Hampton Boulevard, Norfolk, VA, 23529, USA
[d] Department of Strategy, ESC La Rochelle, 102 Rue de Coureilles, 17024, La Rochelle, France

ARTICLE INFO

Keywords:
Queueing theory
Production management
Law of Variability
Serial lines
Throughput

ABSTRACT

Many Queueing Theory and Production Management studies have investigated specific effects of variability on the performance of serial lines since variability has a significant impact on performance. To date, there has been no single summary source of the most relevant research results concerned with variability, particularly as they relate to the need to better understand the 'Law of Variability'. This paper fills this gap and provides readers the foundational knowledge needed to develop intuition and insights on the complexities of stochastic simple serial lines, and serves as a guide to better understand and manage the effects of variability and design factors related to improving serial production line performance, i.e. throughput, inter-departure time and flow time, under random variation.

1. Introduction

A primary concern of any operations management department is to correctly manage production resources to achieve the strategic objectives of the company. This is particularly true for companies gaining competitive advantage through operations. In order to align production performance with the strategic vision of the firm, managers must thoroughly understand and prioritise the most impactful performance factors for factory productivity in order to determine the best course of action to attain the desired goals.

Addressing this need, Schmenner and Swink [1] suggested that the 'Law of Variability' was one of the main laws used by the Operations Management (OM) field to understand the causes of differences in factory productivity, or more generally, factory performance, since variability can have a significant impact on performance; it postulates that '*the greater the random variability, either demanded of the process or inherent in the process itself or in the items processed, the less productive the process is.*'

The most well-known effect of process variability on performance – the variance of inter-arrival and processing (service) times – has been succinctly described by Queueing Theory. A higher variance of inter-arrival and processing times produces longer queues [2] and, consequently, higher mean waiting times for customers and reduced customer satisfaction.

This and other results are captured in 'Factory Physics' by Hopp and Spearman [3]. The tome clearly conveys insights on the fundamental effects of variability on production line performance. The clear insights allow both practitioners and researchers to develop intuition about production line behaviour, a result needed to fully understand the '*consequences of a design decision and the causes of a specific event*' [4].

Other authors have developed comprehensive reviews on Queueing Theory results applied to production lines [5–7] or on the general modelling of stochastic production lines [8,9]. But, most of those efforts focus more on exact modelling of production systems than on supporting the development of intuition about the behaviour of stochastic production lines and the effects of such behaviour on performance, like Hopp and Spearman.

To the best of our knowledge, there is no single resource that compiles the plethora of interesting and relevant results (currently scattered throughout Queueing Theory and Production Management) that could, if assembled, help others gain a better understanding of production lines under the effects of variability. The aim of this paper is to continue the work of Schmenner and Swink [1] and Hopp and Spearman [3] by presenting a summary of the most relevant research

* Corresponding author at: Faculty of Engineering, Universidad Panamericana, Augusto Rodin 498, Insurgentes Mixcoac, 03920, Mexico City, Mexico.
E-mail addresses: rromeros@up.edu.mx, r.romero.silva@hva.nl (R. Romero-Silva), emarsill@odu.edu (E. Marsillac), shaabans@esc-larochelle.fr (S. Shaaban), mhurtado@up.edu.mx (M. Hurtado-Hernández).

https://doi.org/10.1016/j.jmsy.2019.01.005
Received 7 August 2018; Received in revised form 11 December 2018; Accepted 23 January 2019
Available online 13 February 2019

results on the effects of variability in serial production line performance, and provide a reference manual of sorts for practitioners and researchers alike.

We intend that the insights presented here will help readers identify the effects of both diverse variability factors and design features on production line performance. Readers will finish with a better comprehension of the 'Law of Variability' tenets, since the more commonly applied phrase '*random variability*' can represent various factors with varying degrees of impact and does not specifically define any performance measure.

The remainder of this paper is organised as follows: Section 2 presents fundamental definitions and describes the paper's general scope. Section 3 reviews the most relevant results of the effects of variability in the performance of serial production lines. Section 4 presents a brief discussion and Section 5 suggests opportunities for future research. Finally, Section 6 provides conclusions.

2. Scope of the study

The fields of Queueing Theory and Production Management have comprehensively investigated the behaviour of serial production lines under variability.

Most research investigating variability has concentrated on studying the effect of a single variable on the performance of a stochastic production line. Although real production lines are generally more complex, the study of single variable effects can provide great insights under relatively controlled conditions. In addition, intuition can be more easily gained by first studying simpler systems before more realistic but also more complex systems [4]. So, the conclusions here cover studies on simple serial lines or *basic straight lines* [10].

Since our main objective is to assist practitioners and researchers in gaining fundamental understanding, it was decided to concentrate on describing the behaviour of serial lines rather than describing specific formulas for different analytical and approximation models of serial lines. If that subject is of interest, the reader is referred to Buzacott and Shanthikumar [9].

It is worth noting that the intention of this paper is not to create a comprehensive source where all studies on the effects of variability are gathered but rather to create an overview single source where the most relevant conclusions on this topic are collected for an easier understanding of non-expert readers regarding the behaviour of stochastic serial lines.

2.1. Unit of analysis

Serial, stochastic production lines are systems with n number of stations of single resources. They have few constraints and variable production processes, i.e. the processing time of each task is random. *Simple serial lines* (basic straight lines [10]) produce only one type of product, have production and transfer batches of one unit which are manufactured under a First In First Out (FIFO) policy, produce discrete parts, and their operations are unpaced or not synchronised, so each station's operation is independent from the others.

Fig. 1 represents a serial line where S_i is the distribution of processing times in station i, for values of i between 1 and n (n being the total number of stations in the serial line); B_i is the buffer capacity between two consecutive stations; and D is the departure process described by the inter-departure times.

Although most studies consider the same general characteristics of simple serial lines, a major difference between them is the saturation of the line. A saturated serial line is a system where the first station is never starved of work, because as soon as it finishes processing one unit, it starts processing the next without delay. A saturated line is not dependent on any material supply or incoming demand to start processing, and therefore is never constrained by arrivals.

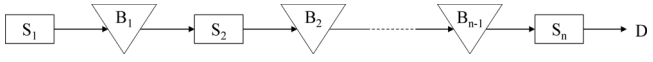

Fig. 1. Graphical representation of a simple serial production line.

Conversely, an unsaturated line is a system that is limited by a stochastic arrival pattern, due to either material supply constraints or customer arrivals. Due to the variability of the arrival and service processes in these systems, the mean system arrival rate (how often customers arrive) needs to be lower than its mean service rate (how quickly customers/orders are processed) to maintain a stable queue [11]. Because of this, the true bottleneck of unsaturated lines is the arrival process. Typical service systems, such as banks, call-centres and hospitals, can be modelled by unsaturated lines, as can production lines in make-to-order environments [12].

Despite these differences, unsaturated lines can be modelled as saturated lines by introducing a *virtual machine* as a first station to model a random arrival process. Since they can both be modelled similarly, both line behaviour types are studied jointly here.

2.2. Performance measures

This subsection defines three complementary measures that have most commonly been used to assess the performance of simple serial lines: throughput, inter-departure times and flow times. These three particular performance measures were selected because the effects of variability and design factors on these measures are, for the most part, straightforward. Other performance measures such as flexibility, quality, cost, and delivery speed and dependability are not covered in this work because that would require studying more complex systems and interactions and would be outside the scope of this study, i.e. simple serial lines.

2.2.1. Throughput

The throughput (TH) rate of a simple serial line is a random variable that measures the actual number of finished product units coming out of the production line per time unit. In Theory of Constraints terms, TH is important for a company because it is the rate at which a factory produces money per time unit. As a random variable, characteristics of the TH distribution can be measured, such as the mean and the variance.

Mean throughput ($\bar{TH}$) for an unsaturated line with infinite buffers and work conserving properties is equal to the mean arrival rate of the system. For a saturated line, $\bar{TH}$ is instead the result of different line parameters interacting with each other (this is further described in Section 3).

While $\bar{TH}$ is commonly considered the main performance measure for saturated lines [13,14], the variance of throughput rate ($Var(TH)$) has also been deemed important because it measures the predictability of the total output per unit time of a factory and the related firm's revenue. Given this value, results for $Var(TH)$ are also included in this paper.

2.2.2. Inter-departure time

The inter-departure time (D) describes the process of two consecutive departures from the production line. The mean inter-departure time ($\bar{D}$) is the inverse of $\bar{TH}$ and is the mean time in which two consecutive departures occur. So, TH can be interpreted as the frequency of production, while D can be interpreted as the period of production. In *Lean* terms, the mean inter-departure time is known as *Takt* time.

Previous studies have been primarily interested in investigating the variance of inter-departure times ($Var(D)$), since it describes the regularity of production output in both saturated and unsaturated lines. The regularity of production output is also relevant for supply chain performance because an irregular output process upstream will affect operations downstream [15].

2.2.3. Flow time

The flow time (*FT*) of an order or customer is the total time the order spent in the production line, either waiting in queue (*W*) or being processed (serviced) by a resource. Flow time (also known as cycle time (*CT*) [16–18]) is substantially related to the quality of service so the effect of variability on flow time and waiting time has been a main topic in unsaturated lines research as well as in assembly line balancing problems [10]. It is worth noting that this paper uses the term *flow time* instead of the term *cycle time* to avoid reader confusion, as this last term can have different interpretations depending on the field of study [19].

$\bar{TH}$ and mean *FT* ($\bar{FT}$) for stochastic production lines with work conserving properties, i.e. lines that don't experience system losses or re-entrant flow, are related by Little's Law [20] through the mean work-in-process ($\bar{WIP}$) of the production line, i.e., $\bar{TH} = \bar{WIP}/\bar{FT}$. The interaction between $\bar{TH}$ and $\bar{FT}$, for a particular production line, is commonly shown in the TH-CT curves [21–24]; however, the effects of different variables on $\bar{TH}$ and $\bar{FT}$ can differ (this is further explained in Section 3).

The study of FT in a single-station system complements the study of *W* in a single station, since $\bar{FT} = \bar{W} + \bar{S}$, where $\bar{W}$ is mean waiting time or queueing time of a single station and $\bar{S}$ is the mean processing time of a single station. Similarly, $\bar{FT}$ for a serial line is equal to the sum of the total waiting times plus the sum of the total processing times of all stations.

For ease of reference, Table 1 shows a summary of all abbreviations used throughout the paper considering the performance measures and different factors included in the study.

3. How different factors affect the performance of production lines under random variation

The 'Law of Variability' posits that the random variability intrinsic to the manufacturing process is detrimental to the productivity of the process. To address this concern, this section first pinpoints how different variability factors inherently associated with the process affect line performance. Second, we relate how some line design factors impact performance. Third, we analyse the effects of some of the most widely known production control techniques on line performance.

3.1. Effects of variability factors

This subsection covers the most widely known effects caused by the variance of inter-arrival and processing times, and the impact of resource unreliability. It also describes several less known factors, e.g., the skewness and autocorrelation of input distributions.

3.1.1. Inter-arrival and processing time variance

Hillier and Boling [25] first showed how a saturated and balanced serial line with variable processing times produced a significantly lower $\bar{TH}$ than what would be expected from a deterministic serial line, i.e. if all the stations of the line have a production rate of one product per time unit, then the mean throughput will be equal to one finished product per time unit. They demonstrated that for saturated lines with $n = 3$, exponential processing times with a mean processing rate equal to 1 and a significantly big buffer size per station equal to 100, $\bar{TH}$ of the production line was less than 1 (i.e. 0.9866), indicating that, even when the buffer size is not a significant constraint on the system, throughput is affected by the processing times' variance. They also showed that by reducing the variance of processing times, overall $\bar{TH}$ increased, without reaching $\bar{TH}$ equal to 1, the value expected for deterministic serial lines.

Later, Conway et al. [26] demonstrated the same behaviour for unbuffered lines with different squared coefficient of variation of service times (SCV_S), being that the squared coefficient of variation of a random variable is the ratio of its variance divided by its squared mean. Serial lines with lower SCV_S produced higher $\bar{TH}$ than serial lines with higher SCV_S having exponential and uniformly distributed processing times. Similarly, Tan [27] suggested that for unbuffered saturated lines, inter-departure time mean and variance increase as SCV_S increases. This type of line behaviour motivates a section on unbalancing saturated serial lines (further discussed in subsection 3.2).

The effects of inter-arrival variance (*Var(A)*) and processing time variance (*Var(S)*) are shown in several of the approximation formulas for $\bar{W}$ and *Var(D)* of an unsaturated single-station line [9]. These formulas show that higher *Var(A)* and/or *Var(S)* result in higher $\bar{FT}$ as well as higher *Var(D)*. The effect of a higher $Var(D_i)$ on an upstream station is a higher $Var(A_{i+1})$ for the next station downstream, since the departure process of an upstream station is the arrival process of the following downstream station. This results in variability propagating throughout an unsaturated serial line, as shown by Wu and Zhao [18].

Taylor and Heragu [28] clearly showed the effect of input variances on $\bar{FT}$ through their investigations on the impact of reduced processing time variance on flow time reduction, compared with a reduction in the mean processing time. They found that a 100% processing times standard deviation improvement equates to reductions of 83% and 67% of mean processing times for infinite and finite buffer sizes, respectively, with (*transformed*) exponential processing times. Likewise, a 100% processing times standard deviation improvement equates to reductions of 22% and 23% of mean processing times for infinite and finite buffer sizes, respectively, with normal processing times.

Khalil et al. [29] also investigated the percentage of blocking (when a machine cannot start processing because the downstream buffer is

Table 1
Summary of abbreviations used throughout the paper.

Factor	Abbreviation	Performance measure	Abbreviation
Mean inter-arrival time	$\bar{A}$	Mean throughput rate	$\bar{TH}$
Variance of inter-arrival time	*Var(A)*	Variance of throughput rate	*Var(TH)*
Squared coefficient of variation of inter-arrival time	SCV_A	Mean inter-departure time	$\bar{D}$
Skewness of inter-arrival time	*Skew(A)*	Variance of inter-departure time	*Var(D)*
Auto-correlation of inter-arrival time	*Corr(A)*	Mean flow time	$\bar{FT}$
Mean service time	$\bar{S}$	Mean Work-In-Process	$\bar{WIP}$
Variance of service time	*Var(S)*	Mean waiting time	$\bar{W}$
Squared coefficient of variation of service time	SCV_S		
Skewness of service time	*Skew(S)*		
Auto-correlation of service time	*Corr(S)*		
Mean resource utilisation percentage	ρ		
Efficiency (reliability) of a production line	ε		
Mean Time to Repair	MTTR		
Mean Time to Failure	MTTF		
Buffer size	*B*		
Line length	*n*		

full) and starving (when a machine cannot start processing because the buffer feeding it is empty) along a balanced saturated serial line with triangular processing times and found that independent of line length, the level of blocking was higher at the first station and lower at the last one. This contrasted to the level of starving, which was lower at the first station and higher at the last one. The highest levels of blocking and starving were also found on the highest values of SCV_S.

3.1.2. Resource unreliability

Perhaps one of the most widely studied factors regarding variability has been the topic of machine unreliability in saturated serial lines. This topic has been commonly studied by modelling resources as Bernoulli machines [30], where the production rate of the system is dependent on the probability of machines' failure at the start of a time period, or by modelling resources as exponential machines with breakdown and repair times modelled as exponential probability distributions. The reader is referred to the work by Li et al. [13] for a comprehensive review of throughput analysis for saturated, unreliable lines with deterministic processing times and finite buffers.

It is worth noting that, even though this line of research has usually considered deterministic processing times [13], their results are included here because of their high relevance for understanding the effect of variability through the adoption of random breakdown events.

The extent of research on Bernoulli and exponential serial lines has proved various system-theoretic properties of serial lines [31] which provide interesting insights about the behaviour of random serial lines. For instance, it has been shown that these lines have the property of *reversibility*, which describes how the $T\bar{H}$ for a saturated serial line with particularly ordered Bernoulli (or exponential) machines and a given set of buffers is equal to the $T\bar{H}$ of an equivalent serial line but with inversely ordered Bernoulli (or exponential) machines and buffers. This property has direct practical implications on the design of serial lines regarding the placement of work and buffer capacity, as will be explained in subsection 3.2.

In addition, it has been proved that $T\bar{H}$ for Bernoulli (or exponential) lines is monotonically increasing on higher machine reliability and increasing buffer capacities. Thus, the *monotonicity* property states that increasing machine reliability (decreasing unreliability) and/or buffer capacity will result in higher throughput.

Li and Meerkov [30] also suggested that production processes with Bernoulli machines should distribute total production work among several machines arranged in series, instead of one single machine capable of doing all the operations since '*longer lines smooth out the production and result in a variability lower than that of one-machine systems*'. This suggests that total unreliability should be distributed throughout a production line instead of being concentrated in just one station.

On the other hand, studies investigating the effects of resource unreliability on saturated production systems with random processing times [32–35] have come to the same conclusions regarding the design of serial lines with unreliable machines. Firstly, lower efficiency (ε) of resource reliability, resulting from the combination of mean machine repair times (MTTR) and mean time-to-failure (MTTF) (ε = MTTF / (MTTF + MTTR)), results in a diminished $T\bar{H}$, which is an equivalent result to the monotonicity property of Bernoulli lines. Secondly, shorter mean repair times with frequent failures are preferred over longer mean repair times with infrequent failures for the $T\bar{H}$ of serial lines with the same average downtimes. This suggests that it is better to carry out frequent, but short, preventive maintenance activities to reduce the likelihood of infrequent, but much longer, reactive repair activities.

3.1.3. Inter-arrival and processing times skewness

Distributions of inter-arrival and processing times can also have higher moments that affect the mean waiting time [36,37]. Some have shown [38,39] that higher inter-arrival skewness (*Skew(A)*) produces lower $\bar{W}$ in unsaturated single queues for different values of the squared coefficient of variation of inter-arrival times (SCV_A) and SCV_S. On the other hand, the effect of processing time skewness (*Skew(S)*) depends on SCV_A values. Atkinson [40] showed that when $SCV_A < 1$ considering a Gamma distribution, $\bar{W}$ increases with increasing processing time (with Erlang distribution) skewness but when $SCV_A > 1$, $\bar{W}$ decreases with increasing processing time skewness.

Lau and Martin [41] suggested that lower values of *Skew(S)* result in higher $T\bar{H}$ in saturated lines. This effect is accentuated in saturated lines when SCV_S is higher and decreased when buffer size and line length increase. They also found that the effect of kurtosis on $T\bar{H}$ is more difficult to interpret since its effect depends on interactions with other factors.

Later, Powell and Pyke [42] observed that negative *Skew(S)* considering the Beta distribution results in higher $T\bar{H}$ in saturated lines. They stated that although the specific effect of kurtosis is difficult to characterise, it can impact $T\bar{H}$ by up to 5%. Moreover, if the modeller only considers the first two moments of the distribution of processing times rather than also the third and fourth, associated with the skewness and kurtosis, Powell and Pyke suggested that an error in $T\bar{H}$ estimation can be as high as 28%.

3.1.4. Auto-correlated processes

Most studies investigating unsaturated lines assume that the arrival process is a renewal process [43], meaning consecutive customer or part arrivals are identically distributed and independent of each other, i.e. not correlated. But, not all unsaturated lines directly receive external demand from customers. For example, a serial production line embedded in a downstream supply chain process usually receives arrivals from an upstream process. Since it has been shown that the departure process of single-resource lines is not a renewal process [44,45] (unless the distribution of inter-arrival and service times is exponential [46]), this suggests the importance auto-correlated arrival effects.

Both Livny et al. [47] and Patuwo et al. [48] show that positively correlated inter-arrival times (*Corr(A)*) increase $\bar{W}$ in a single-station queue with exponential inter-arrival and processing times, and that this effect is higher with increasing values of utilisation. But for some scenarios, negative *Corr(A)* reduce $\bar{W}$ in a single-station queue, as indicated by Nielsen [49] in what he called 'more realistic' auto-correlation patterns.

Livny et al. [47] also show that the effect of autocorrelation on $\bar{W}$ is method dependent, since some of the methods utilised to generate correlated processes [50] exhibited an interaction between the autocorrelation structure of inter-arrival times, processing times and utilisation. For utilisation levels higher than 0.25, most values of positive and negative *Corr(A)* and correlation of processing times (*Corr(S)*) resulted in higher $\bar{W}$.

These conclusions are also shared by Takahashi and Nakamura [51], Altiok and Melamed [52] and Pereira et al. [53], who found that both positive and negative *Corr(A)* and *Corr(S)* have a negative impact on the performance of serial production lines.

3.1.5. Resource utilisation

While resource utilisation (ρ) is not a direct source of variability, but instead, is the result of a combination of random variables (i.e. $\rho = \bar{S}/\bar{A}$, where $\bar{A}$ is the mean inter-arrival time and $\bar{S}$ is the mean service time), it is worth studying since it is subject to random variation and does affect serial line performance.

Resource utilisation affects $F\bar{T}$ for unsaturated lines since higher utilisation of a resource results in higher $F\bar{T}$ [9]. Note that *Var(D)* tends also to be more dependent on *Var(S)* in a single-station line when ρ is high. On the contrary, when ρ is low, *Var(D)* tends to be more dependent on *Var(A)*. This behaviour is shown by the *linking* equation proposed by Hopp and Spearman [3]. It is also interesting to note that when ρ = 1 in an unsaturated single station line with exponential inter-arrival and service times, *Var(TH)* is minimised [54,55]. This effect has been called BRAVO - "Balancing ($\bar{A} = \bar{S}$) Reduces Asymptotic Variance

of Outputs". Similar results have shown the limiting effect of utilisation in the impact of higher moments [56].

Lambrecht and Segaert [57] also showed that buffer content and resource utilisation were not equal for all stations in saturated lines modelled with different processing time distributions, even when all the stations of the line were balanced and no buffer limits were present. On the other hand, Betterton and Silver [58] suggested that *Var(D)* of bottleneck stations should be the lowest of all stations since bottleneck stations are seldom blocked or starved due to their slowest production rate.

3.2. Effects of design factors

3.2.1. Work and variability allocation along the line

One of the most widely accepted conclusions about serial line design is that balanced stochastic serial lines, if comparing two saturated lines with the same assumed $T\bar{H}$, perform worse than unbalanced lines [25,59], which have a *protective capacity* to deal with processing time uncertainty.

These studies generated a field of research on the *bowl phenomenon*, which states that faster stations (non-bottlenecks) of saturated serial lines should be assigned to the middle of the line, while slower stations (bottlenecks) should be positioned at the beginning and end of the line. These and related results can be found in the review papers of Hudson et al. [60] and Mcnamara et al. [14]. Furthermore, equivalent results can be inferred from the *reversibility* property in Bernoulli (or exponential) lines with unreliable machines [31], as the most reliable machine should be placed in the middle of the line to improve $T\bar{H}$.

Regarding station placement, Wu and Zhao [18] provided a heuristic for arranging the stations in an unsaturated line with propagating variability. Suresh and Whitt [61] suggested arranging the stations in a decreasing order of processing time variability, i.e. assigning the station with the lowest SCV_S at the beginning of the line and the station with the highest SCV_S at the end of the line, although the arrangement is not optimal for every configuration. For example, Tembe and Wolff [62] showed the optimal arrangement is to assign the station with the highest mean processing time first in unsaturated serial lines with non-overlapping processing time distributions among the stations [63].

3.2.2. Buffer size and line length

The effects of buffer size (B) and line length (n) under variability have commonly been studied concurrently so their effects are considered jointly here. Although their effects are not the same, they are sometimes considered so by experimental design, e.g., if all the stations have equal B and you increase n, then total buffer capacity for the serial line will increase.

Hillier and Boling [64] and Conway et al. [26] showed that longer saturated lines with uniform and exponential processing times and smaller buffers reduced $T\bar{H}$ in balanced saturated lines, because longer lines and smaller buffers resulted in more station blocking and starving. Tan [27], for Weibull-distributed processing times, also suggested that '*as the number of stations in the line approaches infinity, the cycle time also approaches infinity*' because the longer the line, the lower $T\bar{H}$ and higher *Var(D)*.

Hendricks and Mclain [65] also concluded that as n increases, *Var (D)* increases (considering Uniform, Exponential and Erlang processing times); while as B increases, *Var(D)* decreases. Complementary work by He et al. [66] showed (considering exponentially distributed processing times) that as line length increases, *Var(TH)* decreases but flow time variance increases; while, as B increases, *Var(TH)* increases. Kalir and Sarin [67] present similar conclusions with longer lines with uniform and exponential processing times reducing $T\bar{H}$ and increasing *Var(D)*, and an increase in B increasing $T\bar{H}$ and decreasing *Var(D)*.

When considering the amount of B placed in front of stations in balanced saturated serial lines with limited buffer capacity, Lambrecht and Segaert [57] suggested that if possible, buffers should be placed evenly along the whole serial line. Indeed, if a constraint exists that prevents all stations from having the same buffer size, then buffers should be placed towards the centre of the line, leaving the lowest priority of buffer placement at the start and end of the line. Despite the simplicity of this heuristic, total buffer capacity allocation is an inherently combinatorial problem [68] which complicates optimal performance for all serial line configurations.

As with work placement, the *reversibility* property of Bernoulli (or exponential) serial lines [31] also suggests distributing buffer capacity evenly along the line, or towards the centre of the line if extra units of buffer capacity are available.

In addition, various authors have shown [69–71] that the effect of B on $F\bar{T}$ considering unsaturated lines is similar to the effect on $T\bar{H}$ for the saturated case, as lower values of B will result in a worse (higher) $F\bar{T}$ performance since lower buffer capacity will result in higher probability of blocking along the line.

3.2.3. The complex interactions between variability and design factors

Despite the fact that great intuition can be developed by analysing the effects of various factors in isolation, real serial lines are subject to these factors in an intertwined manner, as Atkinson [40] showed regarding the effects of *Skew(S)*. This means that caution must be exercised since particular factor combinations can produce unexpected results.

For instance, Colledani et al. [72] studied the issue of B in the presence of resource unreliability with geometric distributions and found that *Var(TH)* follows a convex function in relation with B. Either low or high buffer capacity in a two-station line will result in higher *Var (TH)* when compared with an intermediate buffer capacity ($B = 19$). Similar compounding results were presented by Assaf et al. [73], as they suggested that for a two-station line with geometric MTTR and MTTF, the effect of B on *Var(TH)* depends on the value of $T\bar{H}$: if $T\bar{H} < 0.5$, higher B results in higher *Var(TH)*; whereas if $T\bar{H} > 0.5$, higher B results in lower *Var(TH)*, similar to the results from He et al. [66].

Earlier, Tan [74] found that the effect of MTTR on *Var(TH)* in unbuffered balanced lines in the presence of resource exponential unreliability depends on n, when MTTF values remain constant. For example, for two- and three-station lines, higher values of MTTR result in higher *Var(TH)*, but for lines of more than twelve stations, higher values of MTTR result in lower *Var(TH)*. Higher values of MTTR, i.e. lower ε with constant MTTF, resulting in lower *Var(TH)* in longer lines might be explained by the limiting effect of throughput rate values on longer lines with high MTTR values since longer lines with high MTTR values rapidly decrease in overall $T\bar{H}$ values, which imposes a limit on the values of *Var(TH)* caused by increasing values of MTTR. On the other hand, longer lines with lower MTTR values, by having higher $T\bar{H}$ values, have lesser constraints on the possible values that the throughput can take, and therefore, lower MTTR values will result in higher *Var(TH)*. This effect is not present in shorter lines because the throughput is not as constrained as in longer lines, due to the smaller interference among the stations.

Tan [74] also showed that if $\varepsilon \leq 0.5$, higher n results in lower *Var (TH)*; whereas if $\varepsilon > 0.5$, *Var(TH)* is maximised in intermediate values of n ($n > 1$). If ε is close to 1, then higher values of n result in higher *Var (TH)*.

These unreliable lines' results differ from the findings of He et al. [66] on reliable lines. He et al. showed that as B increases, *Var(TH)* monotonically increases and longer lines monotonically decreased *Var (TH)*, which shows that incorporating an added layer of complexity (i.e. unreliability) could produce different performance results.

Another example comes from Hillier's [75] work on the interaction between work and buffer allocation on the profit of a manufacturing firm with exponential and Erlang distributed processing times. In this case, $T\bar{H}$ was considered a revenue-generating factor and inventory was considered a cost. Hillier found that when the cost of inventory

increased, the best pattern regarding work allocation was to assign work towards the start of the line and allocate buffer capacity towards the end of the line, versus following the bowl pattern for work allocation and a balanced pattern for buffer allocation to singularly maximise $T\bar{H}$.

Due to the resulting complexity caused by variability and design factor interactions, some authors [56,66,76] have proposed measuring the overall presence of variability on simple serial lines to assess its impact on the performance of serial lines. Starting with an *idealistic* representation of a serial line where a simple behaviour regarding variability is considered, they compare the idealistic representation with an overall estimate of variability that incorporates a more representative behaviour of variability. Three of these proposed measures are explained in the following section.

3.3. Single measures to assess the impact of variability on the performance

Measuring the variance of critical production line performance measures can provide a good estimate of variability on a production line, but the magnitude of these measures can vary greatly between different production lines, even in the same industrial sector. These measurement differences limit the ability to assess the true impact of variability on performance. For this reason, single measures have been developed to assess the overall impact of variability.

Delp et al. [76] proposed a simple measure to calculate the impact of variability on line performance, and the global impact of resource availability, utilisation and line design. They proposed an '*X-factor*' estimate, by dividing the overall mean flow time of a production line by the sum of all tasks' processing times. The *X-factor* provides a measure to assess the weight of the non-value-adding time of each order or part that remains in the line, in comparison with value-adding operations, i.e. the processing times.

Wu et al. [56] suggested that the variability of a production line (which they termed α) could be estimated by the ratio between the sum of all the waiting times in all the stations and the hypothetical mean waiting time of the bottleneck station considered as an M/M/1 queue. This aligns with conclusions by Suresh and Whitt [77] who noted that production lines with severe bottlenecks and very high utilisation can be reduced to the bottleneck station to calculate the mean waiting times. Accordingly, α describes how much more waiting time is caused by the interactions of different stations and their variability when compared with the waiting time of a single-server queue representing the bottleneck.

Another simple and single measure to assess line variability is the bullwhip effect [78,79], from Supply Chain Management, since it calculates the ratio between the variance of upstream orders (or upstream stations) and the variance of demand (which is fulfilled by downstream inventory). Applied to serial production lines, the measure can assess how variability is dampened or amplified throughout the production line by comparing the variance of the arrival rate against the variance of throughput rate or the variance of inter-arrival times against the variance of inter-departure times.

3.4. Effects of production control techniques

Multiple techniques have proposed dealing with variability in simple serial production lines without modifying line design. These techniques generally focus on the principle of reducing $W\bar{I}P$ in the production line, because increased $W\bar{I}P$ creates longer $F\bar{T}$ for any arriving orders (from Little's Law). $W\bar{I}P$ reduction is normally attained by two different but complementary strategies, 1) setting a production pace, and 2) setting a limit on the $W\bar{I}P$ allowed in the system.

Many popular serial line control techniques use a combination of those two strategies. Kanban [80], CONWIP [81] and DBR [82] methodologies assign a buffer capacity limit to keep $W\bar{I}P$ levels reduced, while maintaining a safety buffer level to cope with uncertainty and avoid station starvation. To further reduce overall $W\bar{I}P$, these methodologies align order releases with inventory consumption since production is only started when consumption is made in different stages of the production line.

Kanban, CONWIP and DBR differ by where they allocate buffer space and by where the order release signal is triggered. While Kanban assigns specific buffer sizes to all stations and then triggers production in each, contingent on the consumption of the next downstream station, CONWIP assigns an overall buffer capacity to minimise starving and blocking and triggers production at the start of the line whenever one part leaves the last line's station. For more details on differences between Kanban and CONWIP and their relation with push and pull production systems, make-to-stock, and make-to-order strategies, refer to Liberopoulos [83].

DBR defines a buffer size for upstream stations, relative to the bottleneck for non-balanced lines, to reduce starving of the bottleneck station, and then sets line pace by triggering production at the first station whenever a part leaves the bottleneck station. Note that buffer allocation in this context is not trivial, since simulation solutions or other methodologies are needed to determine the best buffer configuration for each methodology [84,85].

Other authors have studied the differential performance between popular techniques using theoretical settings of simple serial lines. Lambrecht and Segaert [57] suggested that a DBR-like strategy for pacing the line (i.e. order release) produces better $T\bar{H}$ than a strategy of specifically limiting the buffer on each station. Gstettner and Kuhn [86] suggested that the same $T\bar{H}$ can be attained with smaller overall buffer capacities when using the Kanban system, compared with CONWIP, for saturated lines.

Huang et al. [84] suggested that CONWIP performs better than Kanban and MRP for an unsaturated production line, because it lowers $W\bar{I}P$ and inventory holding costs. Jodlbauer and Huber [87] also found that CONWIP performs better than MRP, Kanban and DBR in terms of $W\bar{I}P$ and service level, but that it is very sensitive to the selection of the correct parameter of total buffer capacity.

For paced saturated lines, Kalir and Sarin [67] found a minimal reduction in both $T\bar{H}$ and *Var(D)* when compared with unpaced saturated lines. Consequently, setting a production pace, although reducing both $W\bar{I}P$ and output variability, might impact $T\bar{H}$ of saturated lines.

3.5. Summary of relevant results

Table 2 provides a summary of serial line factors and performance measures, as well as key authors. Each cell describes how certain factors (rows) influence key performance measures (columns). This reference table can be used by practitioners and researchers to recognise the effect of singular factors on the performance of simple serial lines, when all the other factors remain constant.

In Table 2, "↑" entails an increase in the values of the factor or performance measure, while "↓" entails a decrease; "X" marks whether an increase or decrease in a factor results in an increase or decrease of a singular performance measure.

As reflected in the 'Law of Variability', Table 2 shows that an increased variance (either from inter-arrival or processing times) adversely impacts all performance measures relayed in this paper. In addition, higher resource unreliability clearly influences the performance of serial lines.

More importantly, Table 2 can be viewed as a compendium that lists all the factors that describe *random variability* and how these factors affect line performance when considered in isolation. This then specifies and extends the scope of the 'Law of Variability' for different components. For example, based on Table 2 it could be stated that '*lower processing time skewness results in higher mean throughput rates for saturated lines*' or that '*higher inter-arrival time skewness results in lower mean waiting times for unsaturated lines*'. Table 2 also clarifies how two of the most widely studied design variables regarding serial lines (i.e.

Table 2
Summary of single-factor effects on serial line performance subject to random variation.

		Performance measure							
		$\overline{TH}$ (saturated lines)		*Var(TH)*		*Var(D)*		$\overline{FT}, \overline{W}$	
Factor		↑	↓	↑	↓	↑	↓	↑	↓
Var(A), Var(S)	↑		X [26,64]	X [27]		X [3,9,27]		X [3,9]	
	↓	X [26,64]			X [27]		X [3,9,27]		X [3,9]
ε	↑	X [13,31–35]		See Table 3					
	↓		X [13,31–35]						
MTTF, MTTR	↑		X [32–35] even when ε does not change	See Table 3					
	↓	X [32–35] even when ε does not change							
Skew(A)	↑								X [38,39]
	↓							X [38,39]	
Skew(S)	↑		X [41,42]					See Table 3	
	↓	X [41,42]							
Corr(A)	> 0							X [47,48]	
	< 0							See Table 3	
ρ	↑	X [64]						X [3,9]	
	↓		X [64]						X [3,9]
	=1			minimises *Var(TH)* [54,55]					
B	↑	X [26,27,31, 64,67]		X [66]	See Table 3		X [65,67]		X [69–71]
	↓		X [26,27,31, 64,67]	See Table 3	X [66]	X [65,67]		X [69–71]	
n	↑		X [26,27,30, 64,67]	X [66]	See Table 3	X [27,65,67]		X [9]	
	↓	X [26,27,30, 64,67]		See Table 3	X [66]		X [27,65,67]		X [9]

buffer size and line length) influence performance by showing, for example, that '*shorter serial lines result in higher mean throughput rates as well as lower inter-departure times variance*'.

Despite this ease of reference value, when various variability factors are considered in interaction with each other, conclusions are not as categorical as in the case of *Var(A)*, *Var(S)* or ε. So, it is worth noting that Table 2 only shows a compendium of single relationships between one factor and one performance measure where no interaction with an additional factor has been shown to result in more complex or opposite behaviour. Therefore, to provide with a quick reference for these complex interactions, Table 3 summarises the results from various studies where multiple factors interact to have an effect on a singular performance measure.

While Table 3 serves as a summary reference of complex interactions between multiple factors, caution should be exercised applying Table 3's results, as they are highly dependent on the particular systems' parameters considered in the cited references. Hence, no general conclusions about the effects of these factors can be extended to all simple serial lines, despite the fact that Table 3 provides concise details of the results of the cited references.

4. Discussion

Table 2's summary of the effects of variability and various design factors on different performance measures provides clear and concise guidance to better understand the behaviour of simple serial production lines under the effects of variability. It also contributes to a better understanding of the 'Law of Variability' [1] by describing the effects of different variability factors *inherent to the process*, apart from the variance of inter-arrival and processing times, on the *productivity* of the

Table 3
Summary of multiple-factor effects on the performance of the serial lines studied by the cited references.

Factor 1		Factor 2		Performance measures			
				$\bar{W}$		*Var(TH)*	
				↑	↓	↑	↓
SCV_A	< 1	*Skew(S)*	↑	X [40]			
			↓		X [40]		
	≥ 1	*Skew(S)*	↑		X [40]		
			↓	X [40]			
Corr(A) < 0	$\rho = 0.25$	*Corr(S)*	≤ 0.50		X [47]		
			> 0.50	X [47]			
	$\rho = 0.50$	*Corr(S)*	< 0	X [47]			
			$= 0$		X [47]		
			> 0	X [47]			
	$\rho = 0.80$	*Corr(S)*	$\forall$ *Corr (S)*	X [47]			
B	< 19	*B*	↑			X [72]	
			↓				X [72]
	$= 19$					minimises *Var(TH)* [72]	
	> 19	*B*	↑				X [72]
			↓			X [72]	
$T\bar{T}H$	< 0.5	*B*	↑			X [73]	
			↓				X [73]
	> 0.5	*B*	↑				X [73]
			↓			X [73]	
ε	< 0.5	*n*	↑				X [74]
			↓			X [74]	
	> 0.5	$n > 1$				Value of *n* that maximises *Var(TH)* depends on ε [74]	
	$\rightarrow 1$	*n*	↑			X [74]	
			↓				X [74]
n	≤ 3	MTTR	↑			X [74]	
			↓				X [74]
	> 12	MTTR	↑				X [74]
			↓			X [74]	

process, and how different factors subject to variability influence performance.

A comprehensive understanding of the single effects of variability is important, especially in context of current markets. Today's managers face dynamic market conditions with complex production environments that are difficult to model in an analytical and exact manner. Moreover, Section 3 suggests that firms involved in production line improvement or design processes should be aware of many factors and not only on singular ones when building models of the production line to predict its performance since performance can be gained or lost by different variability and design factors. Building such complicated models that represent the various sources of variation and their impact on performance could also prove difficult to carry out using analytical and approximation models.

This complexity issue might push managers to adopt more intricate modelling tools such as simulation [88–90] to model, analyse and operate production lines. But moving from analytical modelling tools for simple production environments to simulation modelling still requires a solid understanding of and insights into the most important dynamics of stochastic production systems, so as to not inappropriately apply the simulation paradigm [4]. A more comprehensive understanding will enable managers to correctly assess, validate and interpret the results generated by more complex modelling tools, helping them avoid invalid and unreliable results [91–93], and make better decisions with the aim of reaching desired factory performance levels and gain and retain competitive advantage.

This is not to say that all complex systems should be studied and analysed with simulation instead of analytical models, or that simulation models cannot be useful to gain intuition about complex systems. On the contrary, to develop a comprehensive understanding of the intricacies of complex manufacturing and service systems, researchers and practitioners should use every tool at their disposal, as it has been shown that combining analytical and simulation models [94,95] can reap better results when trying to deal with the management of an intricate system.

The specific relevance of the results presented in Table 2 will depend on the reader's objectives and the manufacturing firm's market. For instance, in markets where demand is not a constraint and companies can sell every item produced without high inventory holding costs, firms might be more interested in increasing the mean throughput rate and decreasing throughput variance. Conversely, firms working in environments where demand is lower than production capacity and sets the pace of production might be more concerned with reducing the mean flow/waiting times and maintaining a reduced variance of inter-departure times to provide better customer service.

Clearly, some real environments are inherently more complex than those referenced here, since they produce more than one product or family of products, have additional constraints such as sequence-dependent setup times or delivery dates [96], yield losses, etc. In addition, real environments are concerned with balancing the performance between different, and sometimes contrasting performance measures. A manufacturing firm might be simultaneously concerned with increasing throughput (to increase revenue) while decreasing inventory (to decrease cost). In that situation, addressing a single factor to improve performance might not be the most practical solution since a working combination of multiple factors could produce better results, e.g., CONWIP and DBR by capping buffer levels while setting a production pace equal to demand to reduce flow times.

In spite of these simplifications, we feel that this paper provides the OM field a simple and single referent source that explains the overall behaviour of stochastic serial lines. It can help readers to gain insight into the effects of variability on the performance of manufacturing firms and further understand the implications of the 'Law of Variability'.

Some of the results and papers included are widely known in research fields or covered by previous reviews [5,7,8] and books [6,9,31,97], but the easily interpretable conclusions provided here, have not been collected before in a single reference and represent a valuable contribution to the field. Furthermore, the conclusions provided here can also be applied to the rapidly growing service side of operations, as the operational activities of a number of service-based sectors can also be modelled as serial lines, e.g., healthcare [98,99].

5. Managerial implications and opportunities for future research

For managers, the summaries provided in Tables 2 and 3 and discussed in Section 3.5. provide useful conclusions on the managerial interpretation and application of these results. Complementing those points, some general practical insights can also be inferred. For instance, a common effect for all the performance measurements considered is that both *Var(A)* and *Var(S)* negatively impact the performance of serial lines since increasing *Var(A)* or *Var(S)* or both always affects performance. Thus, efforts to decrease the variance of arrivals and service/processing times will not produce an unexpected outcome in terms of the four performance measurements considered here. A similar conclusion can be inferred from the effects of *B* and *n* on performance, if not for their effects on *Var(TH)*, i.e. higher *B* and lower *n* improve performance for $T\bar{T}H$, $F\bar{T}$ and *Var(D)*, but their effect on *Var (TH)* also depends on other interacting factors, which shows a more complex behaviour.

Overall, it can be said that performance of a serial line is dependent on the interactions among variability factors, e.g., *Var(S)*, and design factors, e.g., *B*, as these interactions create an uneven production flow that results in the starvation or blocking of the stations along the line

and, consequently, a decrease in performance.

Furthermore, performance improvement by increasing B is particularly influenced by the "Law of diminishing returns" [1], as it has been suggested [100,101] that increasing B when B is already very high only results in a marginal performance improvement because the system is already very close to its maximum performance (or performance frontier [1]), which in this case depends on the variance of processing times.

On the other hand, the performance effects of factors such as *Skew (S)*, *Corr(A)* and *Corr(S)* are inherently dependent on the interaction with other factors; thus, performance improvement by modifying these factors could prove to be difficult as it is not as straightforward as modifying performance by improving *Var(A)* or *Var(S)*. These factors also can be difficult to modify since many of the variability reduction techniques have primarily focused on reducing variances [102,103]. Similarly, work, variability and buffer placement are not straightforward tasks since they depend highly on the configuration of the serial line. For example, saturated balanced lines can be improved by either reducing the variance or increasing buffer size towards the middle of the line; but unsaturated balanced lines can show increased performance in some cases by reducing variance or increasing buffer size towards the beginning of the line. Likewise, changing some of the serial line design factors could be difficult since technological constraints might limit the ability to reorder line operators or merge two workstations into a single workstation to reduce line length, for example.

In addition, simple production control techniques, such as, Kanban, CONWIP and DBR, have been found to be great tools to limit the effects of propagating variability along the line by limiting the total amount of work that is allowed into the system. Thus, practitioners could elect to apply these simple control techniques to achieve incremental performance gains before investing in more costly and complex tasks, such as, process improvements (e.g., reduction of MTTF) and production line redesign.

With an improved understanding of the complex interactions summarised here, managers are better able to finesse serial line design factors to enhance performance and deliver results that can support a competitive advantage. Similar to the micro-seconds that separate Olympic winners from mere participants, enhancing the performance of a serial production line, even by small degrees, can generate substantial, compounded results.

Addressing future research directions and opportunities, since most previous studies on simple serial lines focused on the single effects caused by processing time variance in the resulting mean throughput, mean flow time and variance of inter-departure time measures, more efforts are needed to study the impact of specific and various combined factors on particular performance measures.

For instance, most studies investigating the effects of resource unreliability on the performance of serial lines have focused on the effect on $\bar{TH}$ [13] but ignored other equally important performance measures such as those covered in Table 1. Likewise, investigating the factors with highest impact on higher moments than 1 [104] or on the percentiles of the probability distributions of throughput rates and flow times of simple serial lines is rare.

Moreover, because of their analytical tractability, phase-type distributions [105], e.g., exponential, Erlang, and Coxian, have been the most commonly used probability distributions for modelling processing/service times in studies concerned with the effects of variability, as shown in Section 3. However, some authors [103,106–109] have suggested that processing/service times are better modelled by other probability distributions, such as, lognormal and Weibull. Therefore, further research considering more realistic probability distributions is needed as it has been previously shown that queueing network models are highly sensitive to the choice of the probability distribution modelling input distributions [37,39,110,111].

Studies comparing the impacts of different factors on the same line configuration, such as Taylor and Heragu [28] who compared the impact of mean processing time reduction against a reduction in the variance of processing times, could provide interesting insights regarding the leverage of each factor on serial line performance. For instance, studies comparing the impact on performance of a reduction in the variance of processing times against the effect of reducing the variance of time to failure or time to repair could provide very interesting insights regarding the leverage of each factor to improve performance as studies concerned with the impact of unreliability have rarely considered stochastic processing times. Moreover, studies investigating the performance impact of simultaneously modifying two or more factors could provide a better understanding of the combined effects and interactions that the variability factors have on performance, similarly to the results shown by Colledani et al. [72], Atkinson [40] and Johnson [39].

Equivalently, more studies are needed to understand the concurrent impact of different factors on multiple performance measures, as some factors can affect different performance measures in opposing ways. This behaviour can lead managers to find trade-offs between opposing performance effects or to find the best solution regarding an all-encompassing measure of system performance, e.g. the trade-off in terms of profit between the investment and inventory-holding costs of higher buffer capacity and the added revenue resulting from higher throughput created by the extra buffer capacity (see, e.g., [75,112,113]).

In addition, research opportunities exist for the auto-correlated processes impact on various performance measures, since there are relatively few studies that describe the main factors that cause a queue to produce different levels of auto-correlated outputs (see, e.g., [114]). More resource utilisation research is also needed to investigate the various concurrent line performance effects of changing values of utilisation due to varying mean demand rates [115] in serial lines under transient states [116,117].

In terms of measure development, another worthy research direction would be to extend proposed work by Delp et al. [76] and Wu et al. [56], to directly assess the variability of a production line, or more precisely, the impact that the many variability factors (e.g., variance, skewness and auto-correlation of inter-arrival and processing times, and unreliability, etc.) have on the overall performance of the production line. Developing this type of measure would assist assessing if a production line is greatly affected by variability.

A similar reference compilation on more complex production line results, such as assembly lines with merging materials [113,118], setup times [119,120], quality concerns [121], or multiple-product serial lines [122–124] could be useful to better understand the effects of variability in different production environments and further extend the reach of the 'Law of Variability'.

The current market shift from factory-focused, high-volume, single-product serial lines, like the ones studied here, towards customer-responsive [125], multiple-product, complex serial lines with mass customisation [126], highlights the need to further compile conclusions from studies concerned with the impact of variability and design factors on firm's competitive capabilities [127], such as, flexibility, quality, delivery performance and cost, on more complex scenarios. Further efforts to gather and explain these conclusions are particularly needed for a better understanding of current complex systems because companies that want to improve their customer-responsive competitive capabilities can end up affecting other traditional performance measures by increasing the variability and complexity of the system as a result of modifying the design of the manufacturing system (see, e.g., [128,129]).

As was mentioned in Section 4, combining approximation models, exact analytical models, simulation models and even data modelling, e.g., regression analysis and machine learning, will result in both a better understanding and representation of the real manufacturing system in question. Thus, undertaking more studies that integrate different modelling paradigms (see, e.g., [95,130,131]) can provide the

field with a better understanding of the potential of combining different modelling approaches. Despite this, some firms might lack the resources in terms of time, knowledge or capital to build a comprehensive set of models, limiting them to choose, build and use only one modelling approach among the following: approximation, exact analytical, simulation model and data modelling (see, e.g., [132,133]).

Therefore, a stream of research is needed to investigate the trade-off between the estimation accuracy (*Error*) of different modelling paradigms and the time it takes to produce that estimation (*EstimateTime*), depending on system complexity. In this regard, approximation models' estimates can be quickly generated both for simple and complex systems, with the drawback of having a decreasing estimation accuracy as the complexity of the system increases. Exact analytical models are clearly the best models in terms of estimation accuracy; nevertheless, exact analytical models can be difficult to build for very complex systems and the learning curve to use some existing exact analytical models can be very steep, e.g., Matrix-analytic methods [134] and algorithmic analysis based on Markovian decision processes [135].

On the other hand, simulation was seen in previous decades as a time-consuming method that was only the last resort to estimate performance. However, simulation software has significantly evolved in recent years as today moderately complex models can be easily and quickly built with commercial tools, allowing simulation to be a *method of first-resort* [92]. Drawbacks from this approach are the learning curve associated with learning different simulation software, the time to build simulation models from scratch, especially when compared with the time to generate reasonable estimates for simple manufacturing systems with approximation methods, and the reduced accuracy of the simulation models when compared with exact analytical approaches.

A good example of this stream of research can be found in Kim et al. [131], where they compared the performance of simulation and data modelling and described their advantages and limitations, although they did not investigate the difference between the two approaches in terms of *EstimateTime*.

Thus, depending on the complexity and characteristics of the manufacturing system in question, future research can try to determine what is the optimal modelling approach (or combination of approaches) to estimate performance in terms of *Error/EstimationTime*.

Finally, further study could also address the effects and presence of variability in different real manufacturing contexts, and the effects and performance variation of different managerial techniques to cope with variability between sectors, industries or other contextual factors [136–139], based on the constructs of Contingency Theory [140,141]. Integrating context and variability can help practitioners better focus their efforts on managing the effects of variability depending on the manufacturing context. For example, studying the performance of Kanban, CONWIP and DBR in different manufacturing contexts with different degrees of demand uncertainty and process variability could provide novel insights.

6. Conclusions

The main objective and contribution of this paper is to present and summarise some of the most relevant conclusions on the performance behaviour of serial production lines under the effects of variability, and extend the implications of the 'Law of Variability' to improve factory and service management and gain and retain competitive advantage. A brief overview of the most meaningful conclusions on different performance measures is presented and serves as a guide to better understand and manage the effects of variability and design factors on production line performance so managers can exploit the leverage points of factory performance.

This paper fills a gap in literature because few previous efforts have been made to summarise valuable conclusions in a manner which provides practitioners and researchers easier interpretations of the performance effects of various singular variability and design factors. This paper assists readers in developing better understanding and intuition of the behaviour of serial lines so they may design and manage more robust and efficient operations management tasks, before embarking on more complex production systems modelling.

References

[1] Schmenner RW, Swink ML. On theory in operations management. J Oper Manag 1998;17:97–113. https://doi.org/10.1016/S0272-6963(98)00028-X.
[2] Whitt W. The effect of variability in the GI/G/s queue. J Appl Probab 1980;17:1062–71. https://doi.org/10.2307/3213215.
[3] Hopp W. Spearman M. Factory physics. Second. McGraw-Hill; 2000.
[4] Gershwin SB. The future of manufacturing systems engineering. Int J Prod Res 2017;0:1–14. https://doi.org/10.1080/00207543.2017.1395491.
[5] Papadopoulos HT, Heavey C. Queueing theory in manufacturing systems analysis and design: a classification of models for production and transfer lines. Eur J Oper Res 1996;92:1–27. https://doi.org/10.1016/0377-2217(95)00378-9.
[6] Bolch G, Greiner S, de Meer H, Trivedi KS. Queueing networks and markov chains. 1st ed. New York, NY: John Wiley & Sons, Inc.; 1998.
[7] Govil MK, Fu MC. Queueing theory in manufacturing: A survey. J Manuf Syst 1999;18:214–40. https://doi.org/10.1016/S0278-6125(99)80033-8.
[8] Dallery Y, Gershwin S. Manufacturing flow line systems: a review of models and analytical results. Queueing Syst 1992;12:3–94. https://doi.org/10.1007/BF01158636.
[9] Buzacott J, Shanthikumar JG. Stochastic models of manufacturing systems. 1st ed. Englewood Cliffs, New Jersey: Prentice-Hall; 1993.
[10] Battaïa O, Dolgui A. A taxonomy of line balancing problems and their solutionapproaches. Int J Prod Econ 2013;142:259–77. https://doi.org/10.1016/j.ijpe.2012.10.020.
[11] Hillier FS, Lieberman GJ. Introduction to operations research. 9th ed. New York: McGraw-Hill; 2009.
[12] Gupta D, Benjaafar S. Make-to-order, make-to-stock, or delay product differentiation? A common framework for modeling and analysis. IIE Trans 2004;36:529–46. https://doi.org/10.1080/07408170490438519.
[13] Li J, DB E, Huang N, JA M. Throughput analysis of production systems: recent advances and future topics. Int J Prod Res 2009;47:3823–51. https://doi.org/10.1080/00207540701829752.
[14] McNamara T, Shaaban S, Hudson S. Fifty years of the bowl phenomenon. J Manuf Syst 2016;41(1). https://doi.org/10.1016/j.jmsy.2016.07.003.
[15] Hendricks KB. The output processes of serial production lines of exponential machines with finite buffers. Oper Res 1992;40:1139–47. https://doi.org/10.1287/opre.40.6.1139.
[16] Ross SM. Introduction to probability models. 10th ed. Cambridge, Massachusetts: Academic Press; 2009.
[17] Lagershausen S, Tan B. On the exact inter-departure, inter-start, and cycle time distribution of closed queueing networks subject to blocking. IIE Trans 2015;47:673–92. https://doi.org/10.1080/0740817X.2014.982841.
[18] Wu K, Zhao N. Dependence among single stations in series and its applications in productivity improvement. Eur J Oper Res 2015;247:245–58. https://doi.org/10.1016/j.ejor.2015.05.028.
[19] Turpin L. A note on understanding cycle time. Int J Prod Econ 2018;205:113–7. https://doi.org/10.1016/j.ijpe.2018.09.004.
[20] Maxwell WL. Letter to the editor—on the generality of the equation l = λW. Oper Res 1970;18:172–4. https://doi.org/10.1287/opre.18.1.172.
[21] Park S, Fowler JW, Mackulak GT, Keats JB, Carlyle WM. D-optimal sequential experiments for generating a simulation-based cycle time-throughput curve. Oper Res 2002;50:981–90.
[22] Yang F, Ankenman B, Nelson BL. Efficient generation of cycle time-throughput curves through simulation and metamodeling. Nav Res Logist 2007;54:78–93. https://doi.org/10.1002/nav.20188.
[23] Yang F, Ankenman BE, Nelson BL. Estimating cycle time percentile curves for manufacturing systems via simulation. INFORMS J Comput 2008;20:628–43. https://doi.org/10.1287/ijoc.1080.0272.
[24] Sivakumar AI, Chong CS. A simulation based analysis of cycle time distribution, and throughput in semiconductor backend manufacturing. Comput Ind 2001;45:59–78. https://doi.org/10.1016/S0166-3615(01)00081-1.
[25] Hillier FS, Boling RW. The effect of some design factors on the efficiency of production lines with variable element times. J Ind Eng 1966;17:651–8.
[26] Conway R, Maxwell W, McClain JO, Thomas LJ. The role of Work-In-Process inventory in serial production lines. Oper Res 1988;36:229–41.
[27] Tan B. Agile manufacturing and management of variability. Int Trans Oper Res 1998;5:375–88. https://doi.org/10.1111/j.1475-3995.1998.tb00121.x.
[28] Taylor GD, Heragu SS. A comparison of mean reduction versus variance reduction in processing times in flow shops. Int J Prod Res 1999;37:1919–34. https://doi.org/10.1080/002075499190833.
[29] Khalil RA, Stockton DJ, Fresco JA. Predicting the effects of common levels of variability on flow processing systems. Int J Comput Integr Manuf 2008;21:325–36. https://doi.org/10.1080/09511920701233472.
[30] Li J, Meerkov S. Production variability in manufacturing systems: bernoulli reliability case. Ann Oper Res 2000;93:299–324. https://doi.org/10.1023/A:1018928007956.
[31] Li J, Meerkov SM. Production systems engineering. New York, NY: Springer; 2009. https://doi.org/10.1007/978-0-387-75579-3.
[32] Hillier FS, So KC. The effect of machine breakdowns and interstage storage on the

performance of production line systems. Int J Prod Res 1991;29:2043–55. https://doi.org/10.1080/00207549108948066.
[33] Jacobs JH, Etman LFP, Rooda JE. Characterization of operational time variability using effective process times. IEEE Trans Semicond Manuf 2003;16:511–20. https://doi.org/10.1109/TSM.2003.815215.
[34] Schoemig AK. On the corrupting influence of variability in semiconductor manufacturing. Simul. Conf. Proceedings, 1999 Winter, Vol. 1 1999:p. 837–842. https://doi.org/10.1109/WSC.1999.823295.
[35] Patti AL, Watson KJ. Downtime variability: the impact of duration–frequency on the performance of serial production systems. Int J Prod Res 2010;48:5831–41. https://doi.org/10.1080/00207540903280572.
[36] Lemoine AJ. On random walks and stable GI/G/1 queues. Math Methods Oper Res 1976;1:159–64.
[37] Gross D, Juttijudata M. Sensitivity of output performance measures to input distributions in queueing simulation modeling. 1997.
[38] Whitt W. On approximations for queues, III: mixtures of exponential distributions. AT&T Bell Lab Tech J 1984;63:163–75. https://doi.org/10.1002/j.1538-7305.1984.tb00007.x.
[39] Johnson MA. An empirical study of queueing approximations based on phase-type distributions. Commun Stat Model 1993;9:531–61. https://doi.org/10.1080/15326349308807280.
[40] Atkinson JB. Some related paradoxes of queuing theory: new cases and a unifying explanation. J Oper Res Soc 2000;51:921–35. https://doi.org/10.2307/254048.
[41] Lau H-S, Martin GE. The effects of skewness and kurtosis of processing times in unpaced lines. Int J Prod Res 1987;25:1483–92. https://doi.org/10.1080/00207548708919927.
[42] Powell SG, Pyke DF. An empirical investigation of the two-moment approximation for production lines. Int J Prod Res 1994;32:1137–57. https://doi.org/10.1080/00207549408956992.
[43] Mainardi F, Gorenflo R, Vivoli A. Beyond the Poisson renewal process: a tutorial survey. J Comput Appl Math 2007;205:725–35. https://doi.org/10.1016/j.cam.2006.04.060.
[44] Daley DJ. Queueing output processes. Adv Appl Probab 1976;8:395–415. https://doi.org/10.2307/1425911.
[45] Pack C. The output of a D/M/1 queue. SIAM J Appl Math 1977;32:571–87. https://doi.org/10.1137/0132046.
[46] Burke PJ. The output of a queuing system. Oper Res 1956;4:699–704. https://doi.org/10.1287/opre.4.6.699.
[47] Livny M, Melamed B, Tsiolis AK. The impact of autocorrelation on queuing systems. Manage Sci 1993;39:322–39.
[48] Patuwo BE, Disney RL, McNickle CD. The effect of correlated arrivals on queues. IIE Trans 1993;25:105–10. https://doi.org/10.1080/07408179308964296.
[49] Nielsen EH. Autocorrelation in queuing network-type production systems—revisited. Int J Prod Econ 2007;110:138–46. https://doi.org/10.1016/j.ijpe.2007.02.014.
[50] Melamed B. TES: a class of methods for generating autocorrelated uniform variates. ORSA J Comput 1991;3:317–29. https://doi.org/10.1287/ijoc.3.4.317.
[51] Takahashi K, Nakamura N. The effect of autocorrelated demand in JIT production systems. Int J Prod Res 1998;36:1159–76. https://doi.org/10.1080/002075498193264.
[52] Altiok T, Melamed B. The case for modeling correlation in manufacturing systems. IIE Trans 2001;33:779–91. https://doi.org/10.1080/07408170108936872.
[53] Pereira DC, del Rio Vilas D, Monteil NR, Prado RR, del Valle AG. Autocorrelation effects in manufacturing systems performance: a simulation analysis. Proc. Winter Simul. Conf., Winter Simulation Conference. 2012. p. 123:1–123:12.
[54] Nazarathy Y, Weiss G. The asymptotic variance rate of the output process of finite capacity birth-death queues. Queueing Syst 2008;59:135–56. https://doi.org/10.1007/s11134-008-9079-4.
[55] Al-Hanbali A, Mandjes M, Nazarathy Y, Whitt W. The asymptotic variance of departures in critically loaded queues. Adv Appl Probab 2011;43:243–63http://www.jstor.org/stable/23024534.
[56] Wu K, Zhou Y, Zhao N. Variability and the fundamental properties of production lines. Comput Ind Eng 2016;99:364–71. https://doi.org/10.1016/j.cie.2016.04.014.
[57] Lambrecht M, Segaert A. Buffer stock allocation in serial and assembly type of production lines. Int J Oper Prod Manag 1990;10:47–61. https://doi.org/10.1108/01443579010000736.
[58] Betterton CE, Silver SJ. Detecting bottlenecks in serial production lines – a focus on interdeparture time variance. Int J Prod Res 2012;50:4158–74. https://doi.org/10.1080/00207543.2011.596847.
[59] Hillier FS, Boling RW. On the optimal allocation of work in symmetrically unbalanced production line systems with variable operation times. Manage Sci 1979;25:721–8. https://doi.org/10.1287/mnsc.25.8.721.
[60] Hudson S, McNamara T, Shaaban S. Unbalanced lines: where are we now? Int J Prod Res 2015;53:1895–911. https://doi.org/10.1080/00207543.2014.965357.
[61] Suresh S, Whitt W. Arranging queues in series: a simulation experiment. Manage Sci 1990;36:1080–91.
[62] Tembe SV, Wolff RW. The optimal order of service in tandem queues. Oper Res 1974;22:824–32.
[63] Szelki R. Stochastic ordering and dependence in applied probability. 1st ed. New York, NY: Springer; 1995. https://doi.org/10.1007/978-1-4612-2528-7.
[64] Hillier FS, Boling RW. Finite queues in series with exponential or erlang service Times-A numerical approach. Oper Res 1967;15:286–303.
[65] Hendricks KB, McClain JO. The output processes of serial production lines of general machines with finite buffers. Manage Sci 1993;39:1194–201.
[66] He X-F, Wu S, Li Q-L. Production variability of production lines. Int J Prod Econ 2007;107:78–87. https://doi.org/10.1016/j.ijpe.2006.05.014.
[67] Kalir AA, Sarin SC. A method for reducing inter-departure time variability in serial production lines. Int J Prod Econ 2009;120:340–7. https://doi.org/10.1016/j.ijpe.2008.11.016.
[68] Demir L, Tunali S, Eliiyi DT. The state of the art on buffer allocation problem: a comprehensive survey. J Intell Manuf 2014;25:371–92. https://doi.org/10.1007/s10845-012-0687-9.
[69] Perros HG, Altiok T. Approximate analysis of open networks of queues with blocking: tandem configurations. IEEE Trans Softw Eng 1986;12:450–61. https://doi.org/10.1109/TSE.1986.6312886.
[70] Balsamo S, Donatiello L. On the cycle time distribution in a two-stage cyclic network with blocking. IEEE Trans Softw Eng 1989;15:1206–16. https://doi.org/10.1109/TSE.1989.559769.
[71] Wu K, Zhao N. Analysis of dual tandem queues with a finite buffer capacity and non-overlapping service times and subject to breakdowns. IIE Trans 2015;47:1329–41. https://doi.org/10.1080/0740817X.2015.1055389.
[72] Colledani M, Matta A, Tolio T. Analysis of the production variability in multi-stage manufacturing systems. CIRP Ann Manuf Technol 2010;59:449–52. https://doi.org/10.1016/j.cirp.2010.03.142.
[73] Assaf R, Colledani M, Matta A. Analytical evaluation of the output variability in production systems with general Markovian structure. OR Spectr 2014;36:799–835. https://doi.org/10.1007/s00291-013-0343-6.
[74] Tan B. Variance of the throughput of an N-station production line with no intermediate buffers and time dependent failures. Eur J Oper Res 1997;101:560–76. https://doi.org/10.1016/S0377-2217(96)00191-9.
[75] Hillier M. Designing unpaced production lines to optimize throughput and work-in-process inventory. IIE Trans 2013;45:516–27. https://doi.org/10.1080/0740817X.2012.706733.
[76] Delp D, Si J, Fowler JW. The development of the complete X-factor contribution measurement for improving cycle time and cycle time variability. Semicond Manuf IEEE Trans 2006;19:352–62. https://doi.org/10.1109/TSM.2006.879408.
[77] Suresh S, Whitt W. The heavy-traffic bottleneck phenomenon in open queueing networks. Oper Res Lett 1990;9:355–62. https://doi.org/10.1016/01676377(90)90054-9.
[78] Nielsen EH. Small sample uncertainty aspects in relation to bullwhip effect measurement. Int J Prod Econ 2013;146:543–9. https://doi.org/10.1016/j.ijpe.2012.08.004.
[79] Isaksson OHD, Seifert RW. Quantifying the bullwhip effect using two-echelon data: a cross-industry empirical investigation. Int J Prod Econ 2016;171:311–20. https://doi.org/10.1016/j.ijpe.2015.08.027.
[80] Berkley BJ. A review of the KANBAN Production Control Research Literature. Prod Oper Manag 1992;1:393–411. https://doi.org/10.1111/j.19375956.1992.tb00004.x.
[81] Framinan JM, González PL, Ruiz-Usano R. The CONWIP production control system: review and research issues. Prod Plan Control 2003;14:255–65. https://doi.org/10.1080/0953728031000102595.
[82] Betterton CE, Cox JF. Espoused drum-buffer-rope flow control in serial lines: a comparative study of simulation models. Int J Prod Econ 2009;117:66–79. https://doi.org/10.1016/j.ijpe.2008.08.050.
[83] Liberopoulos G. Smith JM, Tan B, editors. Production release control: paced, WIP-based or demand-driven? Revisiting the Push/Pull and make-to-order/make-to-stock distinctions New York, NY: Springer New York; 2013. p. p. 211–47. https://doi.org/10.1007/978-1-4614-6777-9_7.
[84] Huang M, Wang D, Ip WH. A simulation and comparative study of the CONWIP, Kanban and MRP production control systems in a cold rolling plant. Prod Plan Control 1998;9:803–12. https://doi.org/10.1080/095372898233579.
[85] Ye T, Han W. Determination of buffer sizes for drum–buffer–rope (DBR)-controlled production systems. Int J Prod Res 2008;46:2827–44. https://doi.org/10.1080/00207540600922948.
[86] Gstettner S, Kuhn H. Analysis of production control systems kanban and CONWIP. Int J Prod Res 1996;34:3253–73. https://doi.org/10.1080/00207549608905087.
[87] Jodlbauer H, Huber A. Service-level performance of MRP, kanban, CONWIP and DBR due to parameter stability and environmental robustness. Int J Prod Res 2008;46:2179–95. https://doi.org/10.1080/00207540600609297.
[88] Jahangirian M, Eldabi T, Naseer A, Stergioulas LK, Young T. Simulation in manufacturing and business: a review. Eur J Oper Res 2010;203:1–13. https://doi.org/10.1016/j.ejor.2009.06.004.
[89] Negahban A, Smith JS. Simulation for manufacturing system design and operation: literature review and analysis. J Manuf Syst 2014;33:241–61. https://doi.org/10.1016/j.jmsy.2013.12.007.
[90] Pedgen D. Deliver on your promise. 1st ed. Sewickley, PA: Simio LLC; 2017.
[91] Pawlikowski K, Jeong HDJ, Lee JSR. On credibility of simulation studies of telecommunication networks. IEEE Commun Mag 2002;40:132–9. https://doi.org/10.1109/35.978060.
[92] Lucas TW, Kelton WD, Sánchez PJ, Sanchez SM, Anderson BL. Changing the paradigm: simulation, now a method of first resort. Nav Res Logist 2015;62:293–303. https://doi.org/10.1002/nav.21628.
[93] Kelton WD. Methodological expectations for studies using computer simulation. J Bus Logist 2016;37:82–6. https://doi.org/10.1111/jbl.12128.
[94] Alden JM, Burns LD, Costy T, Hutton RD, Jackson CA, Kim DS, et al. General motors increases its production throughput. Interfaces (Providence) 2006;36:6–25. https://doi.org/10.1287/inte.1050.0181.
[95] Lin Z, Matta A, Shanthikumar JG. Combining simulation experiments and analytical models with different area-based accuracy for performance evaluation of manufacturing systems. IISE Trans 2018:1–50. https://doi.org/10.1080/24725854.2018.1490046.

[96] Romero-Silva R, Hurtado M, Santos J. Is the scheduling task context-dependent? A survey investigating the presence of constraints in different manufacturing contexts. Prod Plan Control 2016;27:753–60. https://doi.org/10.1080/09537287.2016.1166274.
[97] Chryssolouris G. Manufacturing systems: theory and practice. 2nd ed. New York, NY: Springer; 2006.
[98] Bhattacharjee P, Ray PK. Patient flow modelling and performance analysis of healthcare delivery processes in hospitals: a review and reflections. Comput Ind Eng 2014;78:299–312. https://doi.org/10.1016/j.cie.2014.04.016.
[99] Zhong X, Lee HK, Williams M, Kraft S, Sleeth J, Welnick R, et al. Workload balancing: staffing ratio analysis for primary care redesign. Flex Serv Manuf J 2016. https://doi.org/10.1007/s10696-016-9258-2.
[100] Gershwin SB, Berman O. Analysis of transfer lines consisting of two unreliable machines with random processing times and finite storage buffers. A I I E Trans 1981;13:2–11. https://doi.org/10.1080/05695558108974530.
[101] Cruz FRB, Van Woensel T, Smith JM. Buffer and throughput trade-offs in M/G/1/K queueing networks: a bi-criteria approach. Int J Prod Econ 2010;125:224–34. https://doi.org/10.1016/j.ijpe.2010.02.017.
[102] Santos J, Wysk R, Torres JM. Improving production with lean thinking. New Jersey, USA: John Wiley & Sons; 2006.
[103] Dudley NA. Work-time distributions. Int J Prod Res 1963;2:137–44. https://doi.org/10.1080/00207546308947819.
[104] Bhat VN. Approximation for the variance of the waiting time in a GI/G/1 queue. Microelectron Reliab 1993;33:1997–2002. https://doi.org/10.1016/0026-2714(93)90356-4.
[105] Latouche G, Ramaswami V. PH distributions. Philadelphia,PA In: Latouche G, Ramaswami V, editors. Society for industrial and applied mathematics1st ed.1999. p. p. 33–60. https://doi.org/10.1137/1.9780898719734.
[106] Inman RR. Empirical evaluation of exponential and independence assumptions in queueing models of manufacturing systems. Prod Oper Manag 1999;8:409–32. https://doi.org/10.1111/j.1937-5956.1999.tb00316.x.
[107] Slack N. Work time distributions in production system modelling. 1982. https://doi.org/10.1002/9781118785317.weom100178.
[108] Brown L, Gans N, Mandelbaum A, Sakov A, Shen H, Zeltyn S, et al. Statistical analysis of a telephone call center: a queueing-science perspective. J Am Stat Assoc 2005;100:36–50.
[109] Armony M, Israelit S, Mandelbaum A, Marmor YN, Tseytlin Y, Yom-Tov GB. On patient flow in hospitals: a data-based queueing-science perspective. Stoch Syst 2015;5:146–94. https://doi.org/10.1214/14-SSY153.
[110] Gross D. Sensitivity of output performance measures to input distribution shape in modeling queues.3. Real data scenario. Simul. Conf. Proceedings, 1999 Winter, Vol. 1 1999:452–7. https://doi.org/10.1109/WSC.1999.823108.
[111] Wu K, Srivathsan S, Shen Y. Three-moment approximation for the mean queue time of a GI/G/1 queue. IISE Trans 2018;50:63–73. https://doi.org/10.1080/24725854.2017.1357216.
[112] Tiacci L. Simultaneous balancing and buffer allocation decisions for the design of mixed-model assembly lines with parallel workstations and stochastic task times. Int J Prod Econ 2015;162:201–15. https://doi.org/10.1016/j.ijpe.2015.01.022.
[113] Romero-Silva R, Shaaban S. Influence of unbalanced operation time means and uneven buffer allocation on unreliable merging assembly line efficiency. Int J Prod Res 2018:1–22. https://doi.org/10.1080/00207543.2018.1495344.
[114] Yeh P-C, Chang J-F. Characterizing the departure process of a single server queue from the embedded Markov renewal process at departures. Queueing Syst 2000;35:381–95. https://doi.org/10.1023/A:1019114732376.
[115] Green LV, Kolesar PJ, Whitt W. Coping with time-varying demand when setting staffing requirements for a service system. Prod Oper Manag 2007;16:13–39. https://doi.org/10.1111/j.1937-5956.2007.tb00164.x.
[116] Yang F, Liu J. Simulation-based transfer function modeling for transient analysis of general queueing systems. Eur J Oper Res 2012;223:150–66. https://doi.org/10.1016/j.ejor.2012.05.040.
[117] Shaaban S, Hudson S. Transient behaviour of unbalanced lines. Flex Serv Manuf J 2012;24:575–602. https://doi.org/10.1007/s10696-011-9110-7.
[118] Sabuncuoglu I, Erel E, Kok AG. Analysis of assembly systems for interdeparture time variability and throughput. IIE Trans 2002;34:23–40. https://doi.org/10.1080/07408170208928847.
[119] Samaddar S. The effect of setup time reduction on its variance. Omega 2001;29:243–7. https://doi.org/10.1016/S0305-0483(00)00045-1.
[120] Kim W, Morrison JR. The throughput rate of serial production lines with deterministic process times and random setups: markovian models and applications to semiconductor manufacturing. Comput Oper Res 2015;53:288–300. https://doi.org/10.1016/j.cor.2014.03.022.
[121] Inman RR, Blumenfeld DE, Huang N, Li J. Designing production systems for quality: research opportunities from an automotive industry perspective. Int J Prod Res 2003;41:1953–71. https://doi.org/10.1080/0020754031000077293.
[122] Kher HV, Fredendall LD. Comparing variance reduction to managing system variance in a job shop. Comput Ind Eng 2004;46:101–20. https://doi.org/10.1016/j.cie.2003.11.002.
[123] Battaïa O, Otto A, Sgarbossa F, Pesch E. Future trends in management and operation of assembly systems: from customized assembly systems to cyber-physical systems. Omega 2018;78:1–4. https://doi.org/10.1016/j.omega.2018.01.010.
[124] Romero-Silva R, Shaaban S, Marsillac E, Hurtado M. Exploiting the characteristics of serial queues to reduce the mean and variance of flow time using combined priority rules. Int J Prod Econ 2018;196:211–25. https://doi.org/10.1016/j.ijpe.2017.11.023.
[125] Schonberger RJ, Brown KA. Missing link in competitive manufacturing research and practice: customer-responsive concurrent production. J Oper Manag 2017;49–51:83–7. https://doi.org/10.1016/j.jom.2016.12.006.
[126] Da Silveira G, Borenstein D, Fogliatto FS. Mass customization: literature review and research directions. Int J Prod Econ 2001;72:1–13. https://doi.org/10.1016/S0925-5273(00)00079-7.
[127] Schoenherr T, Power D, Narasimhan R, Samson D. Competitive capabilities among manufacturing plants in developing, emerging, and industrialized countries: a comparative analysis. Decis Sci 2012;43:37–72. https://doi.org/10.1111/j.15405915.2011.00341.x.
[128] Chryssolouris G, Efthymiou K, Papakostas N, Mourtzis D, Pagoropoulos A. Flexibility and complexity: is it a trade-off? Int J Prod Res 2013;51:6788–802. https://doi.org/10.1080/00207543.2012.761362.
[129] Jönsson M, Andersson C, Ståhl J-E. Relations between volume flexibility and part cost in assembly lines. Robot Comput Integr Manuf 2011;27:669–73. https://doi.org/10.1016/j.rcim.2010.12.002.
[130] Alden JM, Burns LD, Costy T, Hutton RD, Jackson CA, Kim DS, et al. General motors increases its production throughput. Interfaces (Providence) 2006;36:6–25. https://doi.org/10.1287/inte.1050.0181.
[131] Kim BS, Kang BG, Choi SH, Kim TG. Data modeling versus simulation modeling in the big data era: case study of a greenhouse control system. Simulation 2017;93:579–94. https://doi.org/10.1177/0037549717692866.
[132] Priore P, Gómez A, Pino R, Rosillo R. Dynamic scheduling of manufacturing systems using machine learning: an updated review. Artif Intell Eng Des 2014;28:83–97. https://doi.org/10.1017/S0890060413000516.
[133] Srinivas S, Ravindran AR. Optimizing outpatient appointment system using machine learning algorithms and scheduling rules: a prescriptive analytics framework. Expert Syst Appl 2018;102:245–61. https://doi.org/10.1016/j.eswa.2018.02.022.
[134] Latouche G, Ramaswami V. Introduction to matrix analytic methods in stochastic modeling. 1st ed Philadelphia,PA: Society for Industrial and Applied Mathematics; 1999. https://doi.org/10.1137/1.9780898719734.
[135] Tijms HC. Stochastic models: an algorithmic approach. 1st ed. Hoboken, NJ: Wiley; 1995.
[136] Buzacott JA. The structure of manufacturing systems: insights on the impact of variability. Int J Flex Manuf Syst 1999;11:127–46. https://doi.org/10.1023/A:1008050903708.
[137] D'Angelo A, Gastaldi M, Levialdi N. Production variability and shop configuration: an experimental analysis. Int J Prod Econ 2000;68:43–57. https://doi.org/10.1016/S0925-5273(99)00012-2.
[138] Germain R, Claycomb C, Dröge C. Supply chain variability, organizational structure, and performance: the moderating effect of demand unpredictability. J Oper Manag 2008;26:557–70. https://doi.org/10.1016/j.jom.2007.10.002.
[139] CWYCY Wong, Boon-itt S, CWYCY Wong. The contingency effects of environmental uncertainty on the relationship between supply chain integration and operational performance. J Oper Manag 2011;29:604–15. https://doi.org/10.1016/j.jom.2011.01.003.
[140] Sousa R, Voss CA. Contingency research in operations management practices. J Oper Manag 2008;26:697–713. https://doi.org/10.1016/j.jom.2008.06.001.
[141] Romero-Silva R, Santos J, Hurtado M. A note on defining organisational systems for contingency theory in OM. Prod Plan Control Manag Oper 2018;29(16):1343–8. https://doi.org/10.1080/09537287.2018.1535146.